An Optimized framework for Cancer Analysis-an integrated approach of Mining and Evolutionary Algorithms

Research Proposal submitted in partial fulfillment of the requirements for the Degree of

DOCTOR OF PHILOSOPHY
in
ENGINEERING

By
Ashutosh Kumar Dubey

Under the Guidance of
Dr. Umesh Gupta
Associate Professor

JK Lakshmipat University
(Recognized by Govt. of Rajasthan and covered u/s 2(f) of the UGC Act 1956)
Laliya Ka Vas, P.O. Mahapura, Ajmer Road, Jaipur - 302 026 (Rajasthan) India
January, 2012

OUTLINE OF THE PROPOSED TOPIC OF RESEARCH

Name of the Candidate	:	Ashutosh Kumar Dubey
ID No.	:	2012PhDENGG001
Place of Research Work and Organization	:	JKLU, Jaipur
Proposed Supervisor Name	:	Dr. Umesh Gupta
Qualification	:	Ph.D. (Mathematics)
Designation	:	Associate Professor
Organization	:	JKLU, Jaipur

Proposed Topic of Research

An Optimized framework for Cancer Analysis-an integrated approach of Mining and Evolutionary Algorithms

Objectives of the Proposed Research

Many packages have been developed for hospital information systems, and some hospitals have developed their own systems. There are several necessary application areas for a specific hospital information system. Medical data include data about a patient's history, laboratory reports, in-patient data, out-patient data, nursing staff information, specialists' information, other employees' records, surgery schedule, pharmacy information, billing, budget, etc.

The traditional medical record is not only cumbersome but hides much important medical information. However, computerized medical records, containing information about a patient's history, physical examination details, and laboratory findings are more legible and can be made available at several points in the hospital. Computerization of medical records requires some departure from the usual record keeping approach. Moreover, the coding schemes need to be more flexible to allow different doctors to record patient information with different levels of specificity.

Medical data mining has great potential for exploring the hidden patterns in the data sets of the medical domain. These patterns can be utilized for clinical diagnosis. However, the available raw medical data are widely distributed, heterogeneous in nature, and voluminous. These data need to be collected in an organized form. This collected data can be then integrated to form a hospital information system. Data mining technology provides a user-oriented approach to novel and hidden patterns in the data. Data mining and statistics both strive towards discovering patterns and structures in data. Statistics deals with heterogeneous numbers only, whereas data mining deals with heterogeneous fields. We identify a few areas of healthcare where these techniques can be applied to healthcare databases for knowledge discovery. In this dissertation we briefly examine the impact of data mining techniques, including Ant Colony Optimization (ACO), on medical diagnostics. The main objective of the research is to make an efficient cancer detection technique, so that cancer can be detect in the primary stages, so it can be cure.

Background of the Proposed Research

Introduction:

Cancer is a leading cause of death in both developed and developing countries (Jemal A. et. al., 2011). It is estimated that in 2012, there has been over 12.6 million of new cancer cases worldwide, and about 7.6 million cancer deaths. Lung cancer is the most frequent cancer for men, constituting 16.5% of the new cases diagnosed and 22.5% of the total cancer deaths. Breast cancer is the most common cancer in women, accounting for 22.9% of the new cases and 13.7% of the total deaths (Ferlay J. et. al. , 2008 and Liu et. al., 2011). In the US alone in 2010, over 1.5 million people are expected to be diagnosed with cancer, and medical expenditure associated with cancer is estimated to be around 103 billion (Cancer Facts & Figures 2010).

For most cancer types, early diagnosis is regarded as one of the most important determinants for patient survival (Ott J.J. et. al., 2009), and early and precise detection is a critical factor in selecting the proper and effective treatment for the disease (Shah S. et. al.,2007). With the rapid development of information technologies, data mining techniques have been seen applied in various medical applications, and classification, in particular, has been the focus of interest in the field of cancer to help identifying and predicting the disease (Wren Y. Phillips

et. al., 2007). In data mining, classification refers to predicting whether an instance belongs to a certain group using rules discovered from existing data via statistical techniques. Common methods include Artificial Neural Networks, Decision Tree (DT), Support Vector Machine (SVM), Genetic Algorithm (GA), and more recently an Evolutionary Algorithm named Ant Colony Optimization and Particle Swarm Optimization (PSO), of which PSO has demonstrated to competitive in classification tasks, notably in cancer identification and prediction (Sousa T. et. al.,2004).

While in conventional pruning procedure, one attribute is taken out at a time to examine the rule quality (Sousa T. et. al., 2004 & Wang Z. et. al., 2007), for rule which there are multiple limitation conditions in one attribute, the influence of individual parameter inside each attribute is overlooked, and thus it is worth examine each parameter separately.

Analyzing the characteristics of medical data, and a general framework for medical data mining, which consists of medical data selection, medical data pre-processing, medical data mining and knowledge evaluation (Zhao Jitao et. al.,2010) can be useful for disease prediction. We analyze the characteristics of medical dataset and show the framework for data mining in the healthcare and medical domain. For future research, we will apply the framework to medical pattern discovery (Zhao Jitao et. al., 2010).

Applications of data mining technology in medical domain include prediction of the effectiveness of surgical procedures and discovery of relationships among medicine and hospital data. In the last few years, the digital revolution has provided relatively inexpensive and available means to collect and store large amounts of patient data in databases containing rich medical information and made available through the internet for health services globally. Data mining techniques applied on these databases discover relationships and patterns that are helpful in studying the progression of disease. Data mining plays an important role in medical diagnosis. Neither medicine nor medical reasoning represents exact sciences, thus knowledge, which is hidden in patient records is valuable. So Data mining is useful for finding the frequency or occurrences of the symptom from the patient database. By the help of data mining techniques we can find the occurrence, which is used to set the weight of individual item set. Then we may apply Ant Colony Optimization by which we can provide the movement to the symptom in the appropriate group of cancer. An Ant Colony

Optimization (ACO) algorithm is a system consisting of simple agents which cooperate with one another to simulate the behaviour of ants.

The Ant Colony Optimization (ACO) algorithm is a meta-heuristic that has a combination of distributed computation, autocatalysis (positive feedback), and constructive greediness to find an optimal solution for combinatorial optimization problems. This algorithm tries to mimic the ant's behaviour in the real world.

Literature Review of Research Topic

Particle Swarm Optimization (PSO) method for classification of Breast Cancer Data

Tewolde et. al., 2007 exploits the simplicity, efficiency and flexibility of the Particle Swarm Optimization (PSO) method to propose single and multi surface based data separation methods for classification of Breast Cancer Data. The first step of their proposed approaches involve the training of the PSO-based classifiers according to pre-defined data separation methods, using part of the dataset for training. The performances of the classifiers are then tested on the remaining dataset to measure the classification accuracy. The training and testing datasets are derived from the Breast Cancer database obtained from the UCI machine learning repository. Both separation methods produce good classification performance; however, the method based on multiple separating surfaces achieves the best result of 100% classification accuracy on both the training and testing datasets. But the author's does not include any prediction strategy for other cancer diseases.

Particle Swarm Optimization and Support Vector Machine for Gene Signature Selection and Cancer Prediction

Yeung et. al., 2009 improve cancer diagnosis and drug development, the classification of tumor types based on genomic information. As Deoxyribonucleic Acid(DNA) microarray studies produce a large amount of data, expression data are highly redundant and noisy, and most genes are believed to be uninformative with respect to the studied classes. Only a fraction of genes may present distinct profiles for different classes of samples. Classification tools to deal with these issues are thus important. These tools should learn to robustly identify a subset of informative genes embedded in a large dataset that is contaminated with high

dimensional noises. Yeung et. al., 2009 proposes an integrated approach of support vector machine (SVM) and particle swarm optimization (PSO). The proposed approach can simultaneously optimize the selection of feature subset and the classifier through a common solution coding mechanism. As an illustration, the proposed approach is applied to search the combinational gene signatures for predicting histologic response to chemotherapy of osteosarcoma patients. Cross validation results show that the proposed approach outperforms other existing methods in terms of classification accuracy. Further validation using an independent dataset shows misclassification of only one out of fourteen patient samples, suggesting that the selected gene signatures can reflect the chemo resistance in osteosarcoma. Their proposed approach of PSO and support vector machine prediction of chemo-response using an independent dataset of 14 initial biopsy samples achieved 92.9% of accuracy. Authors suggest that further this algorithm is used in similar classification problems with large data sets of nasopharyngeal carcinoma and lung cancer.

Finding Association of Impact Factor for Breast Cancer Patient

Pang et. al., 2009 tries to use a new impact analysis approach to investigating how the age, year and sex affect the probability of getting breast cancer. The impact analysis is an approach that is used to investigate the association of the impact factor with the response variable, and to represent the result in terms of estimated probability. This research considers the breast cancer incident and motility data from the Australian Institute of Health and Welfare since 1983 to 2003. They produce three models in their research. First they observe the effect of sex on breast cancer. Then other two investigations are mainly focus on the impact of different ages and years on breast cancer. Female starts to increase its risk after the age 14, however male breast cancer starts to increase after the age 19. Females' breast cancer incidence rate is 122 times higher than the males. This method is appropriate for evaluating the impact of the risk factors on the incidence of breast cancer. Females' breast cancer incidence rate is 122 times higher than the males.

Gene Association Analysis of Cancer Microarray Data by Pre-processing Algorithms

Baskaran et. al., 2010 Gene association analysis of cancer microarray data provides a wealth of information on gene expression patterns and cancer pathways to enhance the identification of potential biomarkers for cancer diagnosis, prognosis, and prediction of therapeutic

responsiveness. However, achieving these biological/clinical objectives relies heavily on the functional capabilities and accuracy of the various analytical tools to mine these cancer microarray gene expression profiles. Many pre-processing algorithms exist for analysing Affymetrix microarray gene expression data. Previous studies have evaluated these algorithms on their capabilities in accurately determining gene expression using a variety of spike-in as well as experimental data sets. However, variations in detecting differentially expressed genes between these different pre-processing algorithms on a single cancer dataset have not been done in a systems-level evaluation. So (Baskaran et. al., 2010) assessed the comparability and the level of variation between PLIER, GCRMA, RMA and MAS5 for their capability to detect differentially expressed genes.

In terms of consensus, the method had a consensus of 14% and the SAM method 9%, across all 4 pre-processing algorithms. For genes detected by at least 2 or more pre-processing algorithms, the consensus was 46% for the method and 43% for the SAM method. From these analyses, it is unlikely that the gene selection method is a contributing factor to the lack of consensus between genes identified by the different algorithms. They conclude that the sources of variance must reside within the different pre-processing algorithms themselves.

But there is the need of ensemble approach that leverages multiple pre-processing algorithms with cross checking, will offer a better solution in providing a comprehensive view of the true molecular profile of complex biological systems such as cancer.

Optimal syntheses for state constrained problems and optimization of cancer therapies

Piccoli, 2010 resort to sufficient conditions for optimality to achieve results more easy to use. The approach is developed both for candidate value functions and optimal syntheses. Then it is shown on some specific problems in combined cancer therapy.

Optimization of Three Layer Tapered Coaxial Line Applicator for Cancer Thermothrapy by Impedance Matching of Transition Layer Sections

Urata et. al., 2011 presented structural optimization for a coaxial line applicator for the liver cancer thermotherapy. Currently used applicators are slot type antennas made from commercially available semi rigid cables, therefore the impedance matching between an

applicator and a heated liver is not ensured. They presented the new structure by introducing the impedance matching sections between the radiation part and the feeding coaxial line. Scattering parameter S11 is improved from - 9.82 dB to - 18.6 dB at 2.45 GHz by making proper selection of the inserted dielectric materials and optimizing the geometries of matching layers at 2.45 GHz. Prototype applicators are developed and measured with a raw liver of beef and dimethyl sulfoxide (DMSO) by using the vector network analyser (VNA). The measured results reveal that the impedance matching is improved and that DMSO can be used as a phantom material of liver.

Mining Transcriptional Association Rules from Breast Cancer Data

Malpani et. al. , 2011 gain insight into regulatory mechanisms underlying the transcription process of gene expressions, we need to understand the co-expressed gene sets under common regulatory mechanisms. Though computational methods have been developing to identify expression module, challenges still remain for cancer related gene expression profiling. Malpani et. al. , 2011 have developed a method of data preprocessing and two different association rule mining approaches for discovering breast cancer regulatory mechanisms of gene module. Their data preprocessing task involved with two independent data sources: (a) a single breast cancer patient profile data file, (b) a candidate enhancer information data file. Using the integrated data, they also conducted four experiments of the association rule mining.

Future work to enhance the results obtained includes the following Malpani et. al. , 2011:
(1) Currently the p-value is considered to be the only important attribute, and the data is classified and analyzed based on its values. However, position can also play an important role and we can incorporate position value in the future data analysis.
(2) We can make better use of the availability of expression level at different points of time like at 3 hrs, 6 hrs, 12 hrs, 24 hrs, 48 hrs to include various gene expression behavior patterns in the experiment design.

Permittivity Estimation for Breast Cancer Detection Using Particle Swarm Optimization Algorithm

Modiri et. al., 2011 used particle swarm optimization (PSO) algorithm to estimate the permittivities of the tissue layers at microwave frequency band. In addition, breast cancer is an appropriate candidate of MWR due to the breast's exclusive physiology. Several algorithms have been evaluated for analysing the measurement data and solving the inverse scattering problem in MWR, and different levels of accuracy. The potential of PSO in solving this problem is demonstrated at 1-2.25 GHz. Two distinct algorithms are developed for the two considered scenarios. In the first scenario, they assume no a priori knowledge of the tissue under the test, whereas, in the second scenario, a priori knowledge is assumed. However, since these studies under estimate the loss encountered by the test samples, the methods are not valid for body tissue case. It is shown that the algorithm converges relatively fast, and, distinguishes between different tissues with an acceptable accuracy.

Comparative Study of the Optimization Theory for Cancer Treatment

Sbeity Hode et. al., 2011 order to assure hygiene safety for patients, optimal use of drugs is necessary for controlling the number of clinically observed tumor cells. Based on a mathematical model of drug therapy for cancer, and on an objective function, the optimal solution is reached. The model therapy is considered for an individual patient over a fixed time horizon. The disease dynamics are given by an ordinary differential equation that describes the evolution of cancer cells in time. An optimization theory is applied in order to minimize the number of cancer cells at the end of treatment. A comparative study is performed on different treatments profiles applied to different types of cancer. This study shows the usefulness of the suggested mathematical deterministic or hybrid model. It considers the treatment in terms of duration and drug dosage, the initial number of detected cancer cells and the objective function.

Identifying HotSpots in Lung Cancer Data Using Association Rule Mining

Agrawal et. al., 2011 analyse the lung cancer data available from the SEER program with the aim of identifying hotspots using association rule mining techniques. A subset of 13 patient attributes from the SEER data were recently linked with the survival outcome using prediction models, which is used in this study for segmentation. The goal here is to identify characteristics of patient segments where average survival is significantly higher/lower than average survival across the entire dataset. Automated association rule mining techniques

resulted in hundreds of rules, from which many redundant rules were manually removed based on domain knowledge. The resulting rules conform to existing biomedical knowledge and provide interesting insights into lung cancer survival.

Mining Cancer data with Discrete Particle Swarm Optimization and Rule Pruning

Liu et. al., 2011 proposed and implement a classifier using discrete particle swarm optimization (DPSO) with an additional new rule pruning procedure for detecting lung cancer and breast cancer, which are the most common cancer for men and women as per the author's observation. According to the author experiment which shows the new pruning method further improves the classification accuracy and their approach is effective in making cancer prediction. The result shows that for breast cancer data, our proposed DPSO is on par with Naive Bayes with the highest accuracy. And as for the lung cancer data, DPSO performs especially well by getting the highest accuracy of 68.33%, which is around 11.7% and 20.2% better than Naive Bayes and SVM respectively, and 3.8% more accurate than PSO.

Recent Trends in Ant Colony Optimization and Data Clustering

U. Chandrasekhar et. al.,2011 discuss and analyses recent improvements on clustering algorithms like PP (Project Pursuit) based on the ACO algorithm for high dimensional data, recent applications of Data Clustering with ACO, application of Ant-based clustering algorithm for object finding by multiple robots in image processing field and the hybrid PSO/ACO algorithm for better optimized results. According to the author Cluster Analysis is a popular and widely used data analysis and data mining technique. The high quality and fast clustering algorithms play a vital role for users to navigate, effectively organize and structure the data. They observed that Ant Colony Optimization (ACO), a Swarm Intelligence technique, integrated with clustering algorithms, is being used by many applications for past few years.

The Application of Ant Colony Optimization to the Classification Rule Problem

Liang Shyi-Ching et. al., 2011 suggest Classification rule is the most common representation of the rule in data mining. It is based on supervised learning process which generates rules

from training data set. The main goal of the classification rule mining is the prediction of the predefined class based on the group. Based on ACO algorithm, Ant-Miner solved the classification rule problem. According to the author, Ant-Miner shows good performance in many dataset. In this research paper author proposed, an extension of Ant-Miner is proposed to incorporate the concept of parallel processing and grouping. In this paper intercommunication is provided via pheromone among ants is a critical part in ant colony optimization's searching mechanism. The algorithm design in such a way, with a slight modification in this part which removes the parallel searching capability. Based on Ant-Miner, they propose an extension that modifies the algorithm design to incorporate parallel processing. The pheromone trail deposited by ants during the searching procedure affected each other. With the help of pheromone, ants can have better decision making while searching. They provide a possible direction for researches toward the classification rule problem.

Artificial Bee Colony Data Miner (ABC-Miner)

ÇELİK Mete et. al., 2011 discuss about several classical and heuristic algorithms proposed to mine classification rules out of large datasets. In this research authors proposed, a new and novel heuristic classification data mining approach based on artificial bee colony algorithm (ABC) ABC-Miner. Authors proposed approach was compared with Particle Swarm Optimization (PSO) rule classification algorithm and C4.5 algorithm using benchmark datasets. The experimental results show good efficiency of the proposed method.

A Survey of Human Cancer Classification using Micro Array Data

Reena G. Sophia et. al.,2011 suggest that the Cancer research is an interesting research area in the field of medicine. Authors suggest that classification is momentously necessary for cancer diagnosis and treatment. The precise prophecy of dissimilar tumor types has immense value in providing better care and toxicity minimization on the patients. Author suggests that classification of patient taster obtainable as gene expression profiles has become an issue of prevalent study in biomedical research in modern years. Formerly, cancer classification depends upon the morphological and clinical. The modern arrival of the micro array technology has permitted the concurrent observation of thousands of genes, which provoked

the progress in cancer classification using gene expression data. This study hub on the broadly used assorted data mining and machine learning techniques for appropriate gene selection, which pilot to exact cancer classification.

A Cancer Classification Method Based on Association Rules

Wang Gene MeiHua, 2012 propose a novel cancer selection method, AR-SVM. In ARSVM, association rules are used as feature extraction approach to catch the non-linear relation among different genes, and support vector machine is used to classify the transformed gene expression data. Their proposed method achieves both high classification accuracy and good biological interpretability. The experimental results on various gene expression datasets show that AR-SVM achieves the highest classification accuracy in comparison with existing gene expression classification methods.

CAR-Miner: An efficient algorithm for Mining Class-Association Rules

Nguyen et. al., 2013 building a high accuracy classifier for classification is a problem in real applications. One high accuracy classifier used for this purpose is based on association rules. In the past, some researches showed that classification based on association rules (or class-association rules – CARs) has higher accuracy than that of other rule-based methods such as ILA and C4.5. However, mining CARs consumes more time because it mines a complete rule set. Therefore, improving the execution time for mining CARs is one of the main problems with this method that needs to be solved. They propose a new method for mining class-association rule. Firstly, we design a tree structure for the storage frequent item sets of datasets. Some theorems for pruning nodes and computing information in the tree are developed after that, and then, based on the theorems, we propose an efficient algorithm for mining CARs. Experimental results show that our approach is more efficient than those used previously.

Gene selection for cancer tumor detection using a novel memetic algorithm with a multi-view fitness function

Zibakhsh A. et. al., 2013 discuss that Cancer is one of the key research topics in the medical field. An accurate detection of different cancer tumor types has great value in providing

better treatment facilities and risk minimization for patients. Recently, DNA microarray-based gene expression profiles have been employed to correlate the clinical behavior of cancers with the differential gene expression levels in cancerous and benign tumors. An accurate classifier with linguistic interpretability using a small number of relevant genes is beneficial to microarray data analysis and development of inexpensive diagnostic tests. Several well- known and frequently used techniques for designing classifiers from microarray data, such as a support vector machine, neural networks, k-nearest neighbor, and logistic regression model, suffer from low comprehensibility. They propose a new memetic algorithm which is capable of extracting interpretable and accurate fuzzy if–then rules from cancer data. This is the first proposal of memetic algorithms with the Multi-View fitness function approach. The new presented Multi-View fitness function considers two kinds of evaluating procedures. The first procedure, which is located in the main evolutionary structure of the algorithm, evaluates each single fuzzy if–then rule according to the specified rule quality (the evaluating procedure does not consider other rules). However, the second procedure determines the quality of each fuzzy rule according to the whole fuzzy rule set performance. In comparison to classic memetic algorithms, these kinds of memetic algorithms enhance the rule discovery process significantly.

Association rule mining to detect factors which contribute to heart disease in males and females

Nahar Jesmin et. al.,2013 investigates the sick and healthy factors which contribute to heart disease for males and females. Association rule mining, a computational intelligence approach, is used to identify these factors and the UCI Cleveland dataset, a biological database, is considered along with the three rule generation algorithms – Apriori, Predictive Apriori and Tertius. Analyzing the information available on sick and healthy individuals and taking confidence as an indicator, females are seen to have less chance of coronary heart disease then males. Also, the attributes indicating healthy and sick conditions were identified. It is seen that factors such as chest pain being asymptomatic and the presence of exercise-induced angina indicate the likely existence of heart disease for both men and women. However, resting ECG being either normal or hyper and slope being flat are potential high risk factors for women only. For men, on the other hand, only a single rule expressing resting ECG being hyper was shown to be a significant factor. This means, for women, resting ECG status is a key distinct factor for heart disease prediction. Comparingthe healthy status of men

and women, slope being up, number of coloured vessels being zero, and oldpeak being less than or equal to 0.56 indicate a healthy status for both genders.

ACO Sampling: An ant colony optimization-based under sampling method for classifying imbalanced DNA microarray data

Yu Hualong et. al.,2013 propose ACO Sampling that is a novel under sampling method based on the idea of ant colony optimization (ACO) to address this problem. The algorithm starts with feature selection technology to eliminate noisy genes in data. Then they randomly and repeatedly divided the original training set into two groups: training set and validation set. In each division, one modified ACO algorithm as a variant of their previous work is conducted to filter less informative majority samples and search the corresponding optimal training sample subset. At last, the statistical results from all local optimal training sample subsets are given in the form of frequency list, where each frequency indicates the importance of the corresponding majority sample. We only extracted those high frequency ones and combined them with all minority samples to construct the final balanced training set. We evaluated the method on four benchmark skewed DNA microarray datasets by support vector machine (SVM) classifier, showing that the proposed method outperforms many other sampling approaches, which indicates its superiority.

Methodology

The Study

The main purpose of the research study is to explore the previous methodology in the direction of cancer detection and find the advantages and disadvantages so that we can make a better framework where efficient cancer detection technique can be developed, so that cancer can be detect in the primary stages, and it can be cure. After studying several research papers we observe that there is lot of work in this area. But there is still the need of betterment. So our direction of research is to the improvement in terms of cancer detection using Hybrid Technique and develops a java and netbeans based framework which is more efficient in terms of cancer prediction and detection.

The Design

Our Working procedure can be better explained by the following flowchart shown in figure 1.

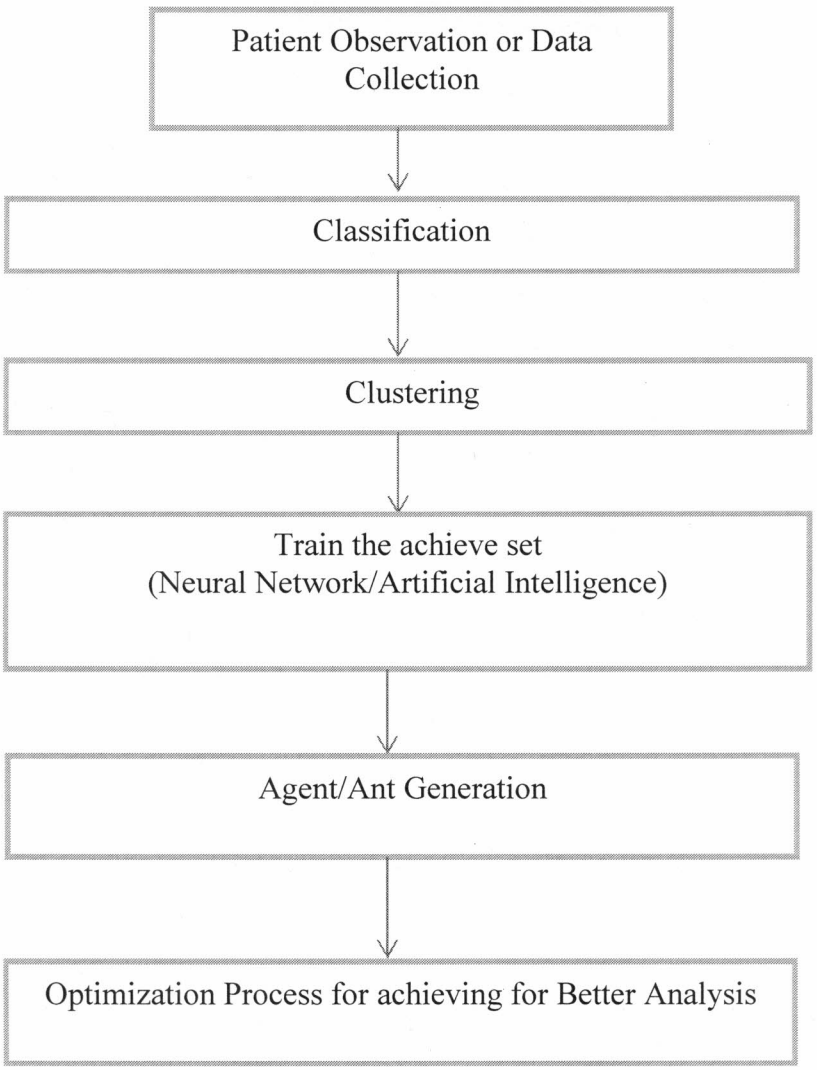


Figure 1: Process Flowchart

In this process we either use the previous research dataset of cancer from UCI machine learning repository or may use the sample patient database for observation. First apply the classification algorithm. Then apply Clustering algorithm for creating clusters of related disease. Then apply training and validation so that it may be achieve a proper trained set and the next observation can trained according to the previous set. Based on the set we generate agents and apply optimization technique to optimize the detection rate.

So our procedure is further divided into six parts:

1) Selection of Data Set

2) Association Rule Classification

3) Clustering

4) Training

5) Agent Generation

6) Optimization Process for achieving Better Analysis

Selection of Data Set (Sample)

We can take the data from UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/datasets.html) or we can take the data from Jawaharlal Nehru Cancer Hospital & Research Centre Bhopal (http://www.jnch.nic.in/).

Association Rule Classification

Although association rules have been predominantly used for data exploration and description, the interest in using them for prediction has rapidly increased in the data mining community. In order to mine only rules that can be used for classification, we can use association rule mining algorithm Apriori to handle user-defined input set constraints. It may classify the rules based on support and confidence.

Frequent Itemset Mining (FIM) is considered to be one of the elemental data mining problems that intends to discover groups of items or values or patterns that co-occur frequently in a dataset. It is of vital significance in a variety of Data Mining tasks that aim to mine interesting patterns from databases, like association rules, correlations, sequences, episodes, classifiers, clusters and the like. The proposed approach utilizes an efficient algorithm called Association Rule Mining Algorithm. The proposed algorithm is employed for the extraction of association rules from the clustered dataset besides performing efficiently when the database consists of very long item sets specifically.

Support count: The support count of an itemset X, denoted by X.count, in a data set T is the number of transactions in T that contain X. Assume T has n transactions (Agrawal, R. et al. , 1993)

$$support = \frac{(X \cup Y).count}{n}$$

.

Confidence: The rule holds in T with confidence conf if conf% of tranactions that contain X also contain Y (Agrawal, R. et al. , 1993)

$$confidence = \frac{(X \cup Y).count}{X.count}$$

Clustering Algorithm

The categorization of objects into various groups or the partitioning of data set into subsets so that the data in each of the subset share a general feature, frequently the proximity with regard to some defined distance measure, is known as Clustering. The clustering problem has been addressed in numerous contexts besides being proven beneficial in many applications. Clustering medical data into small yet meaningful clusters can aid in the discovery of patterns by supporting the extraction of numerous appropriate features from each of the clusters thereby introducing structure into the data and aiding the application of conventional data mining techniques. Numerous methods are available in the literature for clustering. I may employ the renowned K-Means clustering algorithm in our approach.

The k-means algorithm is one of the widely recognized clustering tools that are applied in a variety of scientific and industrial applications. K-means groups the data in accordance with their characteristic values into K distinct clusters. Data categorized into the same cluster have identical feature values. K, the positive integer denoting the number of clusters, needs to be provided in advance.

Training

After finding different clusters training is needed. Training may be applied on the type of cancer detected. It can be possible by applying neural network based training on the obtained clusters. This will be better for the future agents.

Agent Generation

All the trained data which is qualified as a type of cancer symptom can be our agent for the next procedure.

Optimization Process for achieving for Better Detection

All the trained data which is qualified as a type of cancer symptom can be our agent / ant for the next procedure. We may apply Ant Colony Optimization (ACO) algorithm for optimization process.

The Ant Colony Optimization (ACO) algorithm is a meta-heuristic that has a combination of distributed computation, autocatalysis (positive feedback), and constructive greediness to find an optimal solution for combinatorial optimization problems. This algorithm tries to mimic the ant's behavior in the real world.

The ACO algorithm has been inspired by the experiments run (S. Goss et al.,1989) using a colony of real ants. They observed that real ants were able to select the shortest path between their nest and food resource, in the existence of alternate paths between the two

Figure 1 shows the behavior of ants in a double bridge experiment (M. Dorigo et al.,1998). In this case, because of the same pheromone laying mechanism, the shortest branch is most often selected. The first ants to arrive at the food source are those that took the two shortest branches. When these ants start their return trip, more pheromone is present on the short branch than the one on the Long Branch. This behavior was formulated as Ant System (AS)(M. Dorigo et al.,1991). Based on the AS algorithm, the Ant Colony Optimization (ACO) algorithm was proposed(M. Dorigo et al.,1999). In ACO algorithm, the optimization problem is formulated as a graph G = (C; L), where C is the set of components of the problem, and L is the set of possible connections or transitions among the elements of C. The solution is expressed in terms of feasible paths on the graph G, with respect to a set of given constraints. The population of agents (ants) collectively solves the problem under consideration using the graph representation. Though each ant is capable of finding a (probably poor) solution, good quality solutions can emerge as a result of collective interaction amongst ants. Pheromone trails encode a long-term memory about the whole ant search process. Its value depends on the problem representation and the optimization objective.

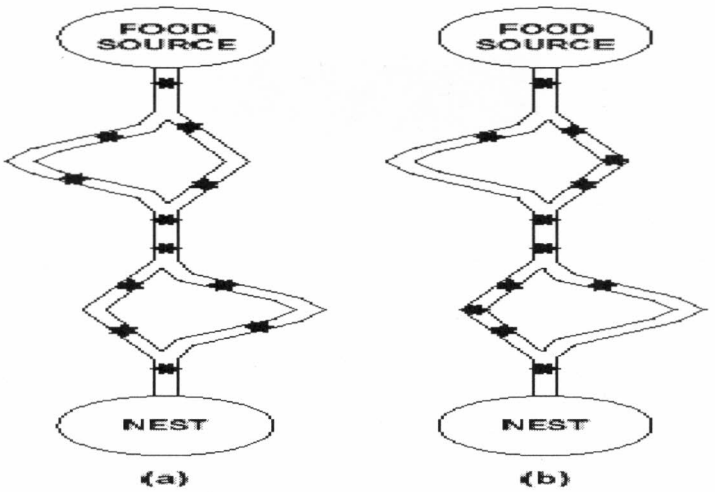
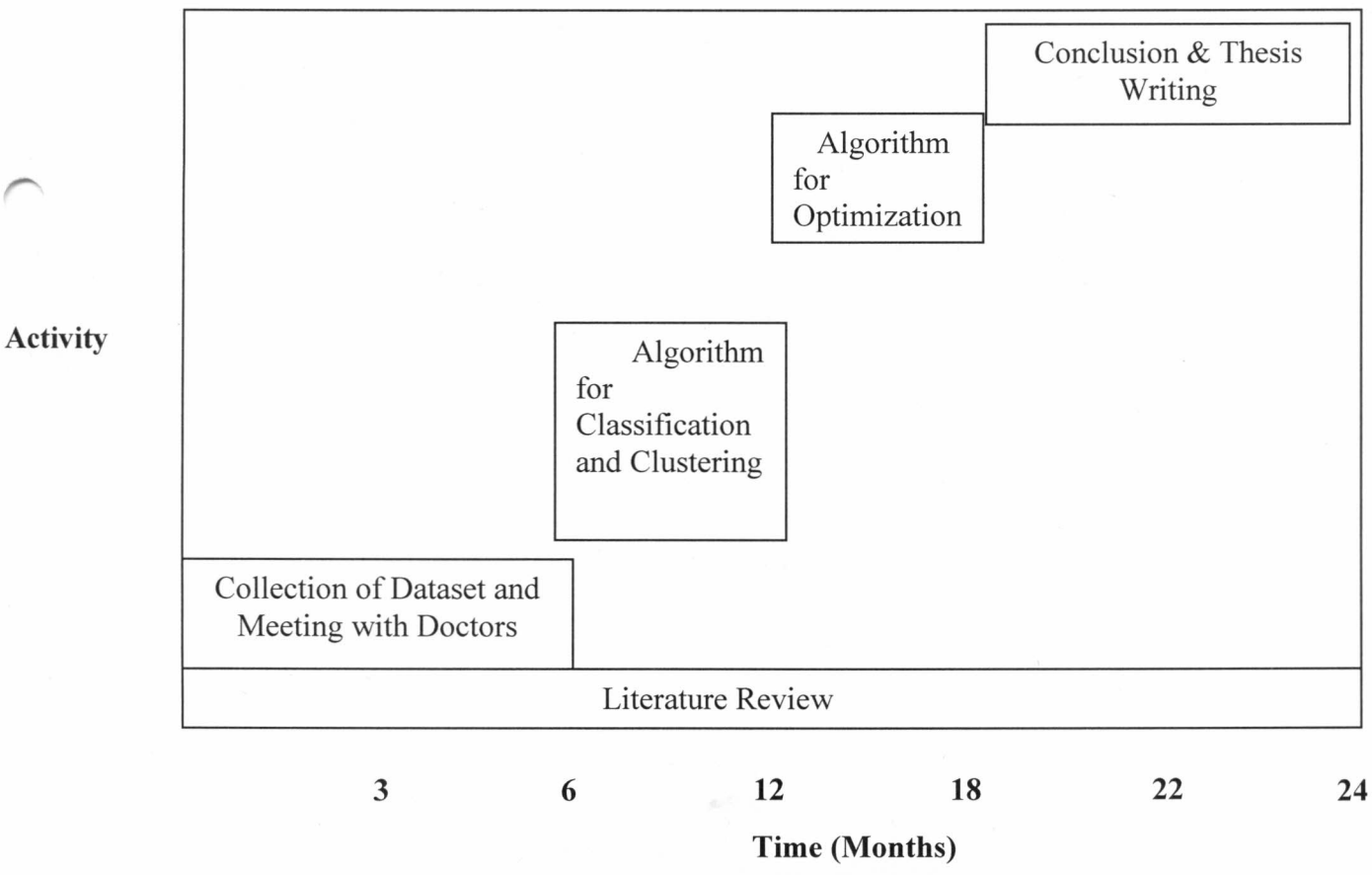


Figure 2: Double bridge experiment. (a) Ants start exploring the double bridge. (b) Eventually most of the ants choose the shortest path (M. Dorigo et al.,1998).

Work Plan

The work-plan for the above-mentioned seven activities is shown in figure 3.

Figure 3: WORK PLAN



REFERENCES

A. Jemal, F. Bray, M.M. Center, J. Ferlay, E. Ward, D. Forman (2011). "Global cancer statistics". CA: a cancer journal for clinicians61 (2): 69–90. doi:10.3322/caac.20107. PMID 21296855.

Agrawal Ankit & Choudhary Alok (2011)." *Identifying HotSpots in Lung Cancer Data Using Association Rule Mining*", IEEE International Conference on Data Mining Workshops, pp. 995-1002.

Agrawal, R., Imielinski, T.,Swami A(1993). *"Mining association rules between sets of items in large databases"*, In Proc. of the ACM SIGMOD Conference on Management of Data, Washington, D.C.,pp. 207-216.

American Cancer Society (2010). *"Cancer Facts & Figures 2010"*.Atlanta: American Cancer Society; 2010.

Baskaran N., Kwoh Chee Keong, Hui Kam M.(2010)." *Outcomes of Gene Association Analysis of Cancer Microarray Data are Impacted by Pre-processing Algorithms*", IEEE International Conference on Bioinformatics and Biomedicine, pp.228-233.

ÇELİK Mete, KARABOĞA Derviş & KÖYLÜ Fehim (2010). *"Artificial Bee Colony Data Miner (ABC-Miner)*", IEEE, pp. 96-100.

J. Ferlay, H.R. Shin, F. Bra, D. Forman, C. Mathers, D.M. Parkin (2008). *"GLOBOCAN 2008 v1.2, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 10"* Lyon, France: International Agency for Research on Cancer.

J.J. Ott, A. Ullrich, A.B. Miller (2009). *"The importance of early symptom recognition in the context of early detection and cancer survival"*, Eur J Cancer 2009;45(16),pp.2743-2748.

Liang Shyi-Ching, Lee Yen-Chun, Lee Pei-Chiang.(2011)*"The Application of Ant Colony Optimization to the Classification Rule Problem"*, IEEE International Conference on Granular Computing , pp. 390-392.

Liu Yao & Chung Yuk Ying(2011).*"Mining Cancer data with Discrete Particle Swarm Optimization and Rule Pruning"*, pp. 31-34.

M. Dorigo,M. Maniezzo,A. Colorni(1991*)."The Ant Systems: An Autocatalytic Optimizing Process"*, Revised 91-016, Dept. of Electronica, Milan Polytechnic.

M. Dorigo, Gianni Di Caro, Luca M. Gambardella(1998)."*Ant Algorithms for Discrete Optimization. Technical Report Tech. Rep. IRIDIA/98-10, IRIDIA"*, Universite Libre de Bruxelles, Brussels, Belgium.

M. Dorigo & G. Di Caro(1999). New Ideas in Optimisation. McGraw Hill, London, UK.

Malpani Rakhi, Lu Meiliu, Zhang Du, Sung Wing Kin(2011).*"Mining Transcriptional Association Rules from Breast Cancer Profile Data"*, IEEE IRI, pp. 154-159.

Modiri Arezoo & Kiasaleh Kamran(2011)." *Permittivity Estimation for Breast Cancer Detection Using Particle Swarm Optimization Algorithm"*, International Conference of the IEEE EMBS Boston, Massachusetts USA, pp. 1359-1362.

Nahar Jesmin, Imam Tasadduq, Tickle Kevin S., Chen Yi-Ping Phoebe(2013)." *Association rule mining to detect factors which contribute to heart disease in males and females"*, Elsevier, Expert Systems with Applications 40 (2013),pp. 1086–1093.

Nguyen Loan T.T., Vo Bay, Hong Tzung-Pei, Thanh Hoang Chi(2013)." *CAR-Miner: An efficient algorithm for mining class-association rules"*, Elsevier, Expert Systems with Applications 40(2013), pp.2305–2311.

Pang, Kwok-Pan & Ali,A B M Shawkat(2009).*"Finding Association of Impact Factor for Breast Cancer Patient- A Novel Statistical Approach"*, Second International Conference on Environmental and Computer Science, pp. 449-453.

Piccoli Benedetto (2010). *"Optimal syntheses for state constrained problems and optimization of cancer therapies"*, IEEE Conference on Decision and Control, Hilton Atlanta Hotel, Atlanta, GA, USA, pp. 7033-7038.

Reena G. Sophia and Rajeswari P.(2011). *"A Survey of Human Cancer Classification using Micro Array Data"*, IJCTA, Vol 2 (5), pp. 1523-1533.

S. Goss, S. Aron, J. L. Deneubourg, J. M. Pasteels (1989). *"Self-organized Shorcuts in the Argentine Ant"*, Naturwissenschaften, Vol. 76, pp.579–581.

Sbeity Hode, Younes Rafic, Topsu Suat, Mougharbel Imad. *"Comparative Study of the Optimization Theory for Cancer Treatment"*, 4th International Conference on Biomedical Engineering and Informatics (BMEI), pp. 927-933.

S. Shah, A. Kusiak (2007). *"Cancer gene search with data-mining and genetic algorithms"*, Computers in Biology and Medicine, 37(2), pp.251–261.

T. Sousa, A. Silva, A. Neves (2004). *"Particle swarm based data mining algorithms for classification tasks"*, Parallel Computing, 30(5–6), pp. 767–783.

Tewolde, Girma S. & Hanna, Darrin M. (2007). *" Particle Swarm Optimization for Classification of Breast Cancer Data using Single and Multisurface Methods of Data Separation"*. IEEE EIT Proceedings, pp. 443-446.

U. Chandrasekhar, P. Naga Poojitha Rao. *" Recent Trends in Ant Colony Optimization and Data Clustering: A Brief Survey"*, IEEE ,pp. 32-36.

Urata Masako, Kimura Satoshi, Wakino Kikuo, Kitazawa Toshihide(2011). *"Optimization of Three Layer Tapered Coaxial Line Applicator for Cancer Thermothrapy by Impedance Matching of Transition Layer Sections"*, IEEE, pp. 120-123.

Wang MeiHua, Su XiongBin, Liu FuMing, Cai RuiChu. *" A Cancer Classification Method Based on Association Rules"*,9th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2012),pp. 1094-1098.

Y. Phillips-Wren, G. Sharkey (2007). *"Mining lung cancer patient data to assess healthcare resource utilization"*. Expert Systems with Applications.

Yeung, C.W. & Chan, K.Y.(2009*)." An Integrated Approach of Particle Swarm Optimization and Support Vector Machine for Gene Signature Selection and Cancer Prediction"*. Proceedings of International Joint Conference on Neural Networks, Atlanta, Georgia, USA, pp. 3450-3456.

Yu Hualong, Ni Jun, Zhao Jing (2013)." *ACO Sampling: An ant colony optimization based under sampling method for classifying imbalanced DNA microarray data"*, Elsevier, Neuro computing 101(2013), pp.309–318.

Z. Wang, X. Sun, D. Zhang (2007). "*A PSO-Based Classification Rule Mining Algorithm*", Lecture Notes in Computer Science, Vol.4682, pp. 377–384,.

Zhao Jitao & Wang Ting(2010).*"A General Framework for Medical Data Mining"*, International Conference on Future Information Technology and Management Engineering, pp. 163-165.

Zibakhsh A., Abadeh M.Saniee(2013)." *Gene selection for cancer tumor detection using a novel memetic algorithm with a multi-view fitness function"*, Elsevier, Engineering Applications of Artificial Intelligence 26(2013), pp.1274–1281.